HG-Dagger: Interactive Imitation Learning with Human Experts

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, Mykel J. Kochenderfer

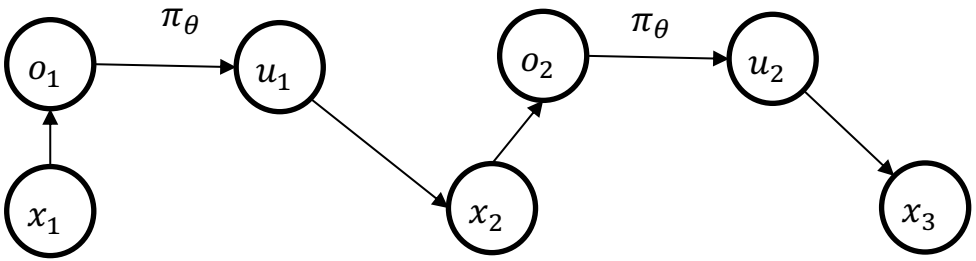
Presenter: Kun Qian Instructor: Yuke Zhu

Nov. 2nd. 2021

Motivation: Imitation Learning

- Motivation: The agent needs to learn a policy whose resulting states, action trajectory distribution matches the expert's trajectory distribution

Challenges: dependent actions and states



Imitation Learning: Background

What to Learn?

1. Direct Feedback from humans: Using the human feedbacks in different states as labels to train the model; Need to address the issue of action interdependence; (This paper)
2. Learn a reward function: Try to infer the latent rewards/goals of the teacher(inverse RL)

Who is the teacher?

1. Human (This paper)
2. Optimal Controller

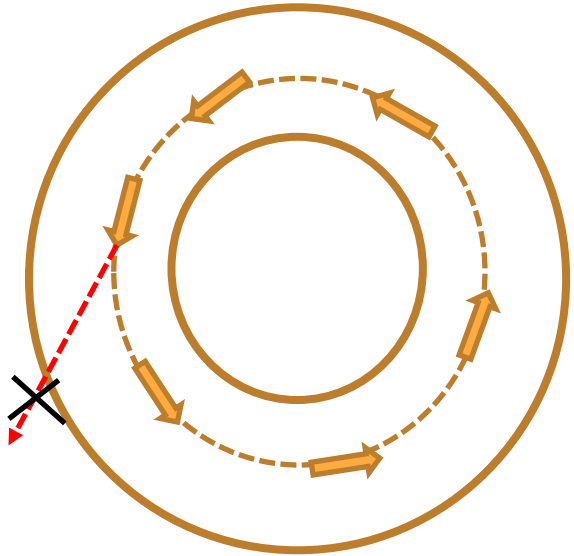
Imitation learning as supervised learning

- Learn from the data: Let the agent's policy π_θ learn from human's policy π^*

Basic version: Behavior Cloning, directly learn from human data $\mathcal{D}_{\pi^*} = \{o_t, u_t\}, 1 \leq t \leq N$

Assumptions: actions & trajectories are i.i.d

Issue:
1. Accumulating error
2. Data distribution mismatch



Dagger as a solution

Dagger steps

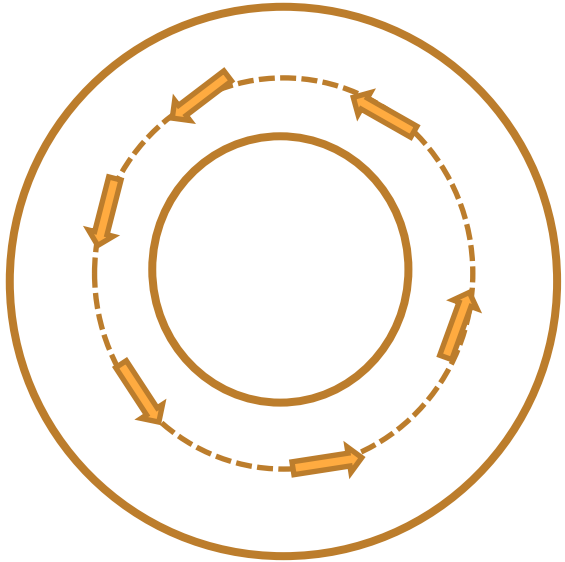
1. Train $\pi_\theta^t(u_t|o_t)$ from human data $\mathcal{D}_{\pi^*} = \{o_1, u_1, \dots, o_N, u_N\}$

2. Run $\beta_t \pi^* + (1 - \beta_t)\pi_\theta^t(u_t|o_t)$ to get a dataset $\mathcal{D}_\pi = \{o_1, \dots, o_M\}$

3. Human provides labels $\{u_1, \dots, u_M\}$

4. $\mathcal{D}_{\pi^*} \leftarrow \mathcal{D}_{\pi^*} \cup \mathcal{D}_\pi$

5. Repeat the above steps



Dagger's problem

1. The switch between human and robot may cause disability and the behavior of human may be influenced;

2. Tricky to control the parameter β_t ;

3. Robot decides when to ask for human demonstration;

4. Safety issue: Executing an imperfect policy will cause accidents in real world experiments;

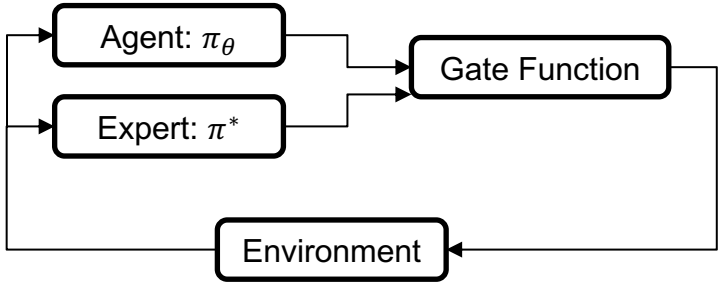
Ensemble Dagger

Motivation: To find a good indicator to decide if the agent should ask for human demonstration

Methods: Use an ensemble of neural networks trained from the data to represent the policy, add a gate function to decide the involvement of human demo

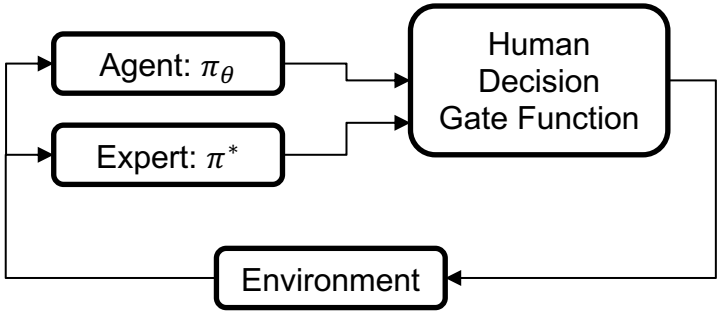
Gate function: if $||\overline{u_\theta} - \mathrm{u}^*|| \geq \hat{\tau}$, or $\sigma_{u_\theta} \geq \hat{\sigma}$, we ask for human demo

Remaining problem: How to find a good threshold?



HG-Dagger Algorithm

Motivation: Give human more control over the sampling and demonstration process



HG-Dagger steps

1. Train $\pi_\theta^t(u_t|o_t)$ from human data $\mathcal{D}_{\pi^*}^0 = \{o_1, u_1, \ldots, o_N, u_N\}$

2. Run $\pi_\theta^t(u_t|o_t)$;If human decides to take control, then record state and control into $\mathcal{D}_\pi^t = \{o_1^t, u_1^t \ldots\}$

3. $\mathcal{D}_{\pi^*}^{t+1} \leftarrow \mathcal{D}_{\pi^*}^t \cup \mathcal{D}_\pi^t$

4. Repeat the above steps

HG-Dagger Algorithm

Motivation: To better analyze the algorithm, the concept “policy Confidence” is adopted

Ensemble of Neural Networks: The agent/novice’s policy is described by an ensemble of neural networks

Definition of doubt: $d_N(o_t) = \ || \ diag(C_t)||_2$

Safe Threshold: $\tau = \frac{1}{len(\mathcal{I})/4} \sum_{t=0.75N}^{N} (\mathcal{I}[t])$, $\mathcal{I}[t]$ includes the doubt level when human interrupts. This means this paper chooses the threshold according to the later stage demonstrations

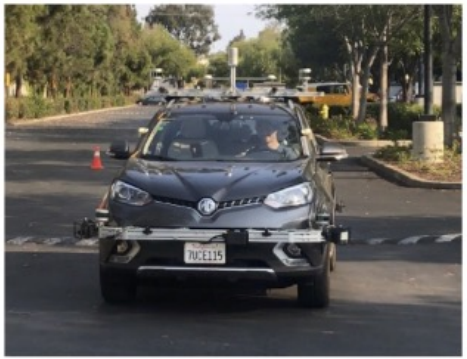
Experimental Setup

General setting: Teach an agent to drive along a two-lane single direction road with static cars present as obstacles

Inputs to agent: distance from median, orientation, speed, distance to edge of current lane, distance to nearest obstacle

Actions: wheel steering angle, vehicle speed

Metric: road departure time and collision rate



Experimental Setup

Training data:

Initialization: 10,000 (states, action) pairs

Additional Demonstration: 2,000 demonstrations

Baseline models:

Model 1: Behavior Cloning

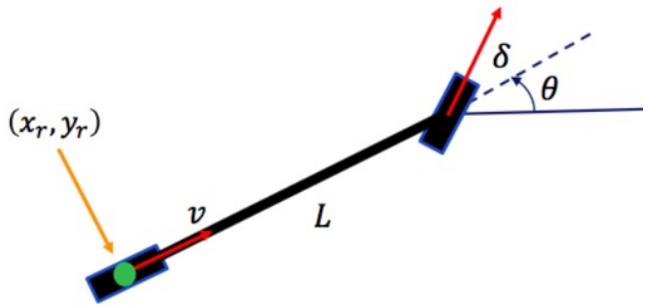
Model 2: Dagger, ($\beta_t = 0.85^t$)

Experiment details

Real-world experiment: Human steer wheel to take control

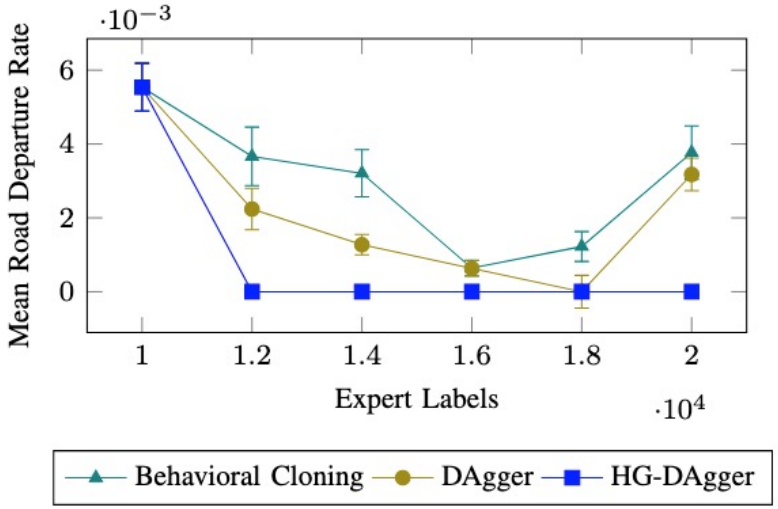
Simulation: Bicycle model

Bicycle Model



$$\dot{x} = v\cos(\theta)$$
$$\dot{y} = v\sin(\theta)$$
$$\dot{\theta} = v\tan(\delta)/L$$
$$\dot{\delta} = u$$

Experimental Results: Simulation



Interesting phenomenon: Dagger deteriorate in later epochs;

Analysis: smaller beta, larger algorithm control; caused by imperfections/mismatch between train and test data distribution

Experimental Results: Simulation

TABLE I: Mean collision and road departure rates per meter, and mean road departure duration in seconds, for rollouts initialized within or outside the permissible set.

Initialization	Collision Rate	Road Departure Rate	Departure Duration
$\hat{\mathcal{P}}$	0.607×10^{-3}	0.607×10^{-3}	1.630
$\hat{\mathcal{P}}'$	7.533×10^{-3}	12.092×10^{-3}	3.740

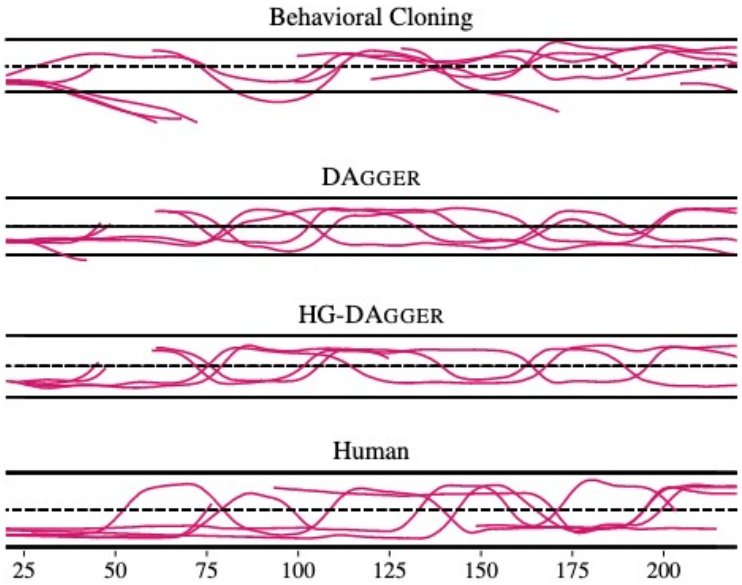
We construct another set S

$$S = \{(x, y, \theta, s) \mid y \in [-6, 6] \text{ meters}, \theta \in [-15, 15] \text{ degrees}, s \in [4, 5] \text{ m/s}, \max(d_l, d_r) < 8 \text{ meters}\}$$

$\hat{\mathcal{P}} = \{x_t | d_N(\mathcal{O}(x_t)) \leq \tau\}$, $\hat{\mathcal{P}}'$ is the complement of $\hat{\mathcal{P}}$

The initializations are uniformly sampled from $\hat{\mathcal{P}} \cap S$ and $\hat{\mathcal{P}}' \cap S$

Experimental Results: Real-World



Bhattacharyya distance:

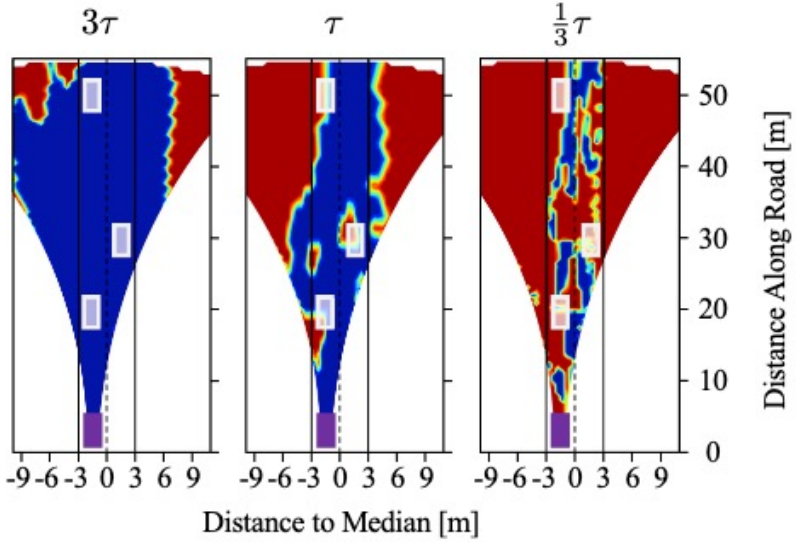
$$D_B(p,q) = -\ln\big(BC(p,q)\big), BC(p,q) = \sum_x \sqrt{p(x)q(x)}$$

Compare the steering angle distribution according to Bhattacharyya distance

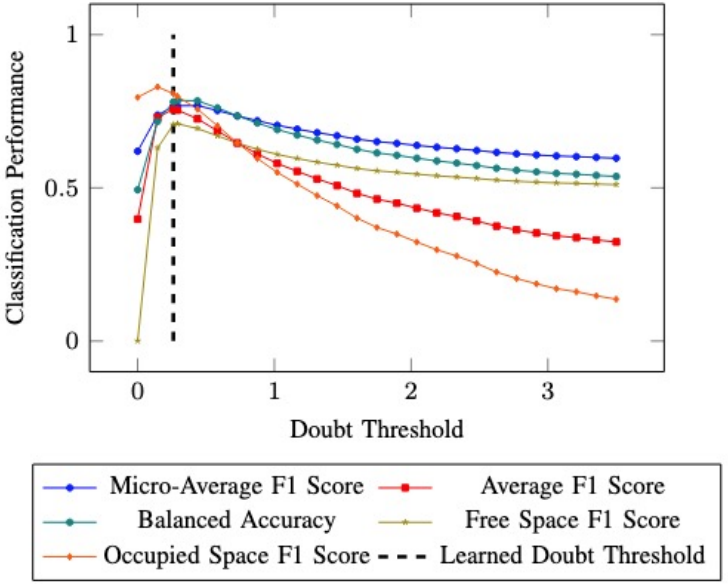
TABLE II: Summary of on-vehicle test data. Totals are for the first 5,000 samples collected.

	# Collisions	Collisions Rate	# Road Departures	Road Departure Rate	Bhattacharyya Metric
Behavioral Cloning	1	0.973×10^{-3}	6	5.837×10^{-3}	0.1173
DAGGER	1	1.020×10^{-3}	1	1.020×10^{-3}	0.1057
Human-Gated DAGGER	**0**	**0.0**	**0**	**0.0**	**0.0834**

Experimental Results: Real-World



The heat map to valid the selection of confidence threshold. Heat map is drawn according to binary classification with different confidence threshold



Quantitative analysis of the binary classification performance. Free space -- $\hat{\mathcal{P}}$; Occupied Space -- $\hat{\mathcal{P}}'$

Experiment Conclusions

1. HG-Dagger outperforms Dagger in both simulation and real-world experiments in terms of collision rate and out-of-road rate
2. The confidence threshold derived from human judgement is shown to be reasonable
3. Dagger is shown to have poor performance in later stages

Open Issues

1. Comparison with other methods are in lack
2. Large scale real-world experiments to verify performance
3. Not really safe when testing: no human interruption in test, only interruption in training

Future Directions

1. Connect the doubt-based risk metric with execution risk
2. Use doubt-based metric as a gating function to switch between different sub-policies

Extended Readings

1. Menda K, Driggs-Campbell K, Kochenderfer MJ. Ensembledagger: A bayesian approach to safe imitation learning. In2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2019 Nov 3 (pp. 5041-5048). IEEE.
2. Ross S, Gordon G, Bagnell D. A reduction of imitation learning and structured prediction to no-regret online learning. In Proceedings of the fourteenth international conference on artificial intelligence and statistics 2011 Jun 14 (pp. 627-635). JMLR Workshop and Conference Proceedings.
3. Saunders W, Sastry G, Stuhlmueller A, Evans O. Trial without error: Towards safe reinforcement learning via human intervention. arXiv preprint arXiv:1707.05173. 2017 Jul 17.

Summary

1. Introduced the imitation learning problem;
2. Went through the Dagger algorithm and list the existing problems of Dagger;
3. Introduced HG-Dagger algorithm in detail;
4. Analyzed the Experiment results of HG-Dagger;
5. Analyzed the open issues and possible future directions in safe imitation learning;